New Statistical Techniques for Predictive Water Quality Modeling at Great Lakes Beaches

ATER QUALITY
OWCAST: POOR
owcast" system is being tested on this
ach to predict bacterial levels that
may be present in the water.
OOR WATER QUALITY
PREDICTED TODAY
on conditions observed this morning. This
s that bacteria levels are likely to be high.
imming is not advised, especially for
ren, the elderly, and those in ill health.
ody water contact may result in illness.
LLECTED BY:
District
NOWCAST SUPPORT PROVIDED BY:
Cleveland Lakefront State Park
Cleveland Department of Public Health
Cuyahoga County Board of Health
Ohio Department of Health
United States Geological Survey
ation, call (216) 432-7301
onowcast.info

Michael N. Fienen[1]
Wesley R. Brooks[1]
Steven R. Corsi[1]
Kurt Wolfe[2]
Rajbir Parmar[2]
Mike Galvin[2]
Mike Cyterski[2]

[1]*USGS Wisconsin Water Science Center, Middleton, WI*
[2]*USEPA Office of Research and Development, Athens, GA*

Introduction

E. coli is used as a fecal indicator bacterium

Without rapid methods, 24 hours are required to get an answer

Environmental predictors (weather, waves, etc.) are available in real time

Our ultimate goal is to predict *E. coli* from environmental predictors.

NO INGIERA AGUA DEL LAGO.
NADE A SU PROPIO RIESGO.
Para mas información:
www.wibeaches.us
TSIS TXHOB HAUS COV DEJ NO.
TXHOB DA DEJ NTAWM NO.
Xav paub ntau tshaj no:
www.wibeaches.us
NO HAY AVISO PUESTO EN ESTE MOMENTO.
TSIS MUAJ YAM TXAUS NTSHAI TAM SI NO.
PRECAUCION: AVISO DE CALIDAD DE AGUA. AUMENTO DE RIESGO DE ENFERMEDAD PROBABLE BASADO EN RECIENTES ANALISIS DE LA BACTERIA E. COLI.
XYUASZOO: DEJ QIAS HEEV TSAM TAU KAB MOB RAWS LI TSHAWB POM KAB HMOB E. COLI.
PLAYA CERRADA. NO ENTRE AL AGUA. GRAN RIESGO DE ENFERMEDAD PROBABLE.
NTUG DEJ KAW. TXHOB NKAG MUS TOM HAV DEJ TSAM TAU KAB MOB.

Modeling Approaches

getting the most information from all data available

persistence model

using yesterday's *E. coli* to predict today's conditions

ordinary least squares regression (OLS)

well-established technique

requires decisions on how to handle interaction and correlation

partial least squares (PLS)

newer technique used extensively in spectroscopy

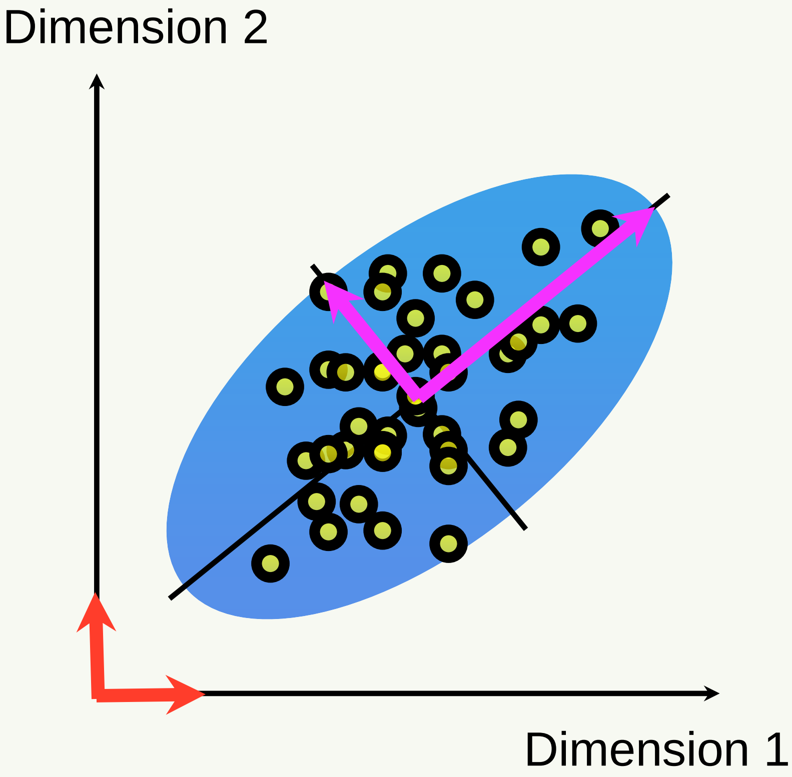
introduced to beach community by Hou, Rabinovici, and Boehm (2006)

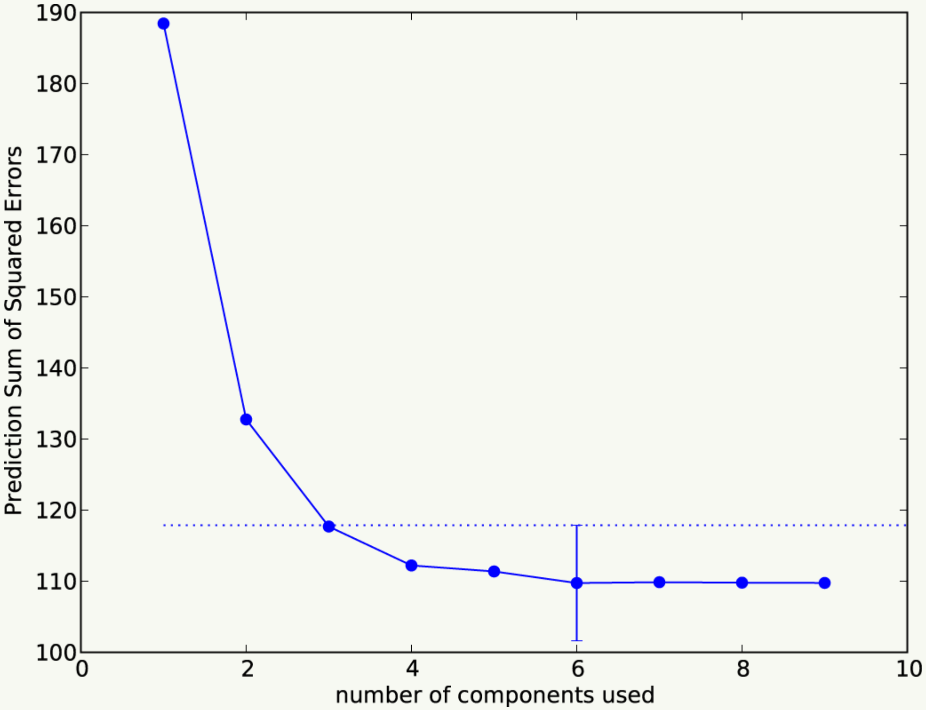
useful when variables are correlated or insensitive

overfitting prevented through cross validation and component trimming

algorithm replaces trial-and-error interaction terms & variable selection

Partial Least Squares: Example





Partial Least Squares

regression built on principal directions relating variables to predictions

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

error term

coefficients

E. coli observations

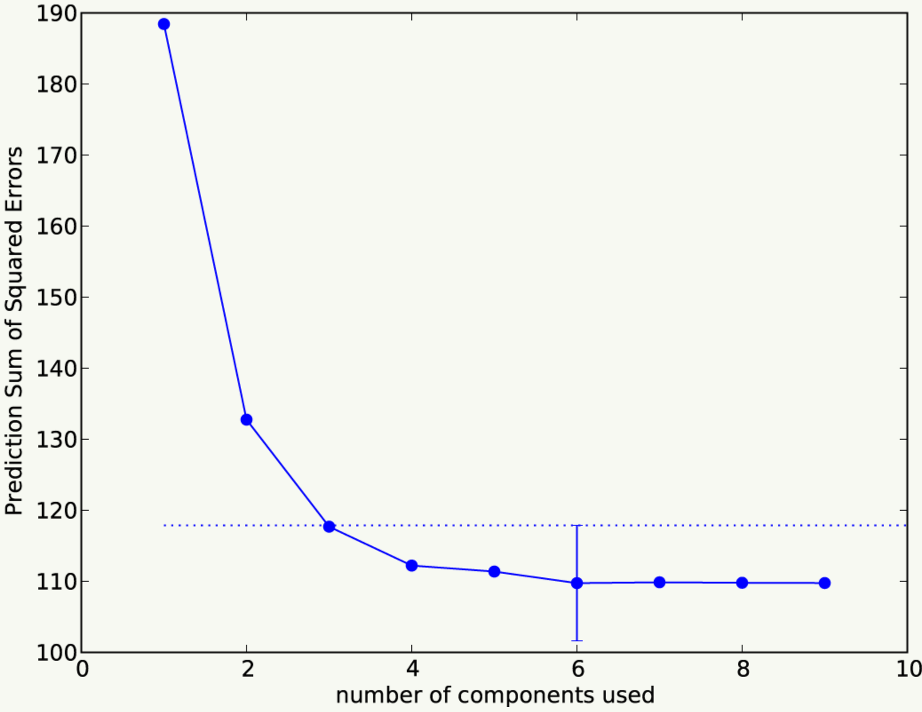
predictor variables

principal directions of covariance between $\mathbf{X}$ and $\mathbf{y}$ define components

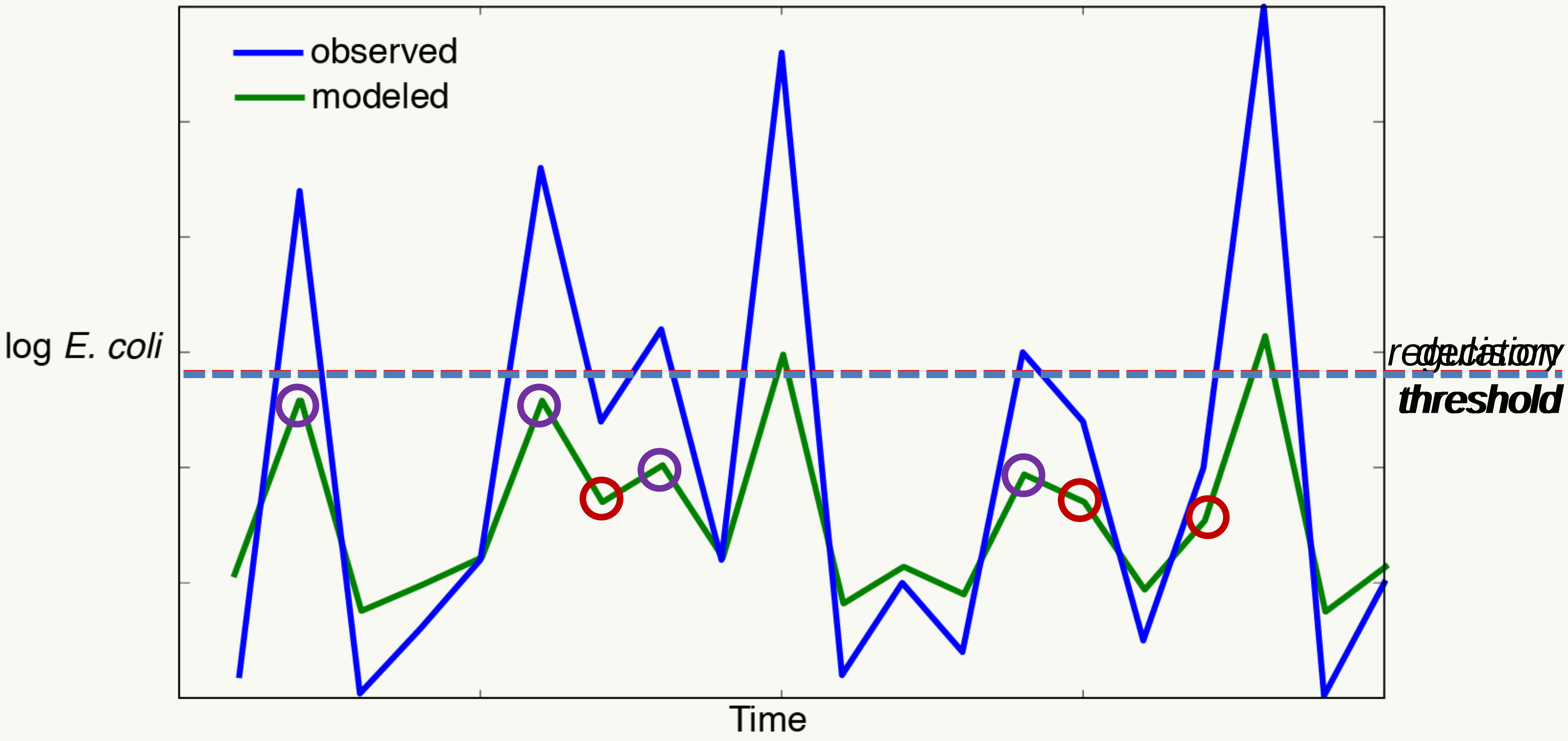
using all components is equivalent to OLS with all variables

the number of components retained is chosen balancing lower PRESS against overfitting

each component includes information from all base variables



Decision and Regulatory Thresholds

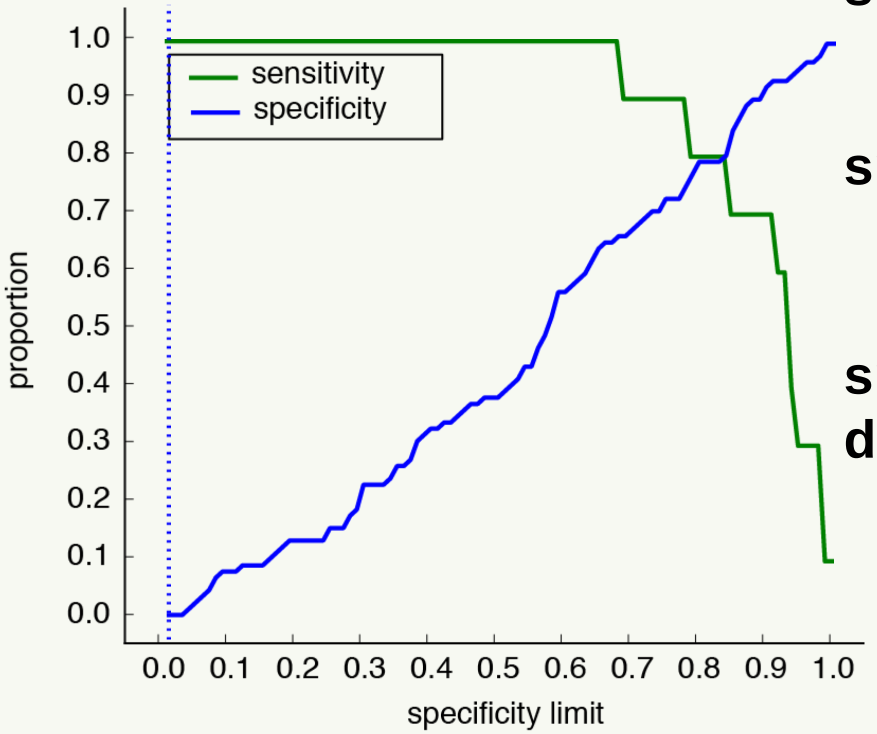


Tradeoffs for Model Performance

robust methods and improved data integration

building a PLS model

managers decide the tradeoff between protective and permissive



sensitivity: proportion of true positives

high sensitivity means increasing

true positives

specificity: proportion of true negatives

low specificity means increasing

false positives

specificity limit (vertical blue line**) is the dial controlling this tradeoff**

USGS

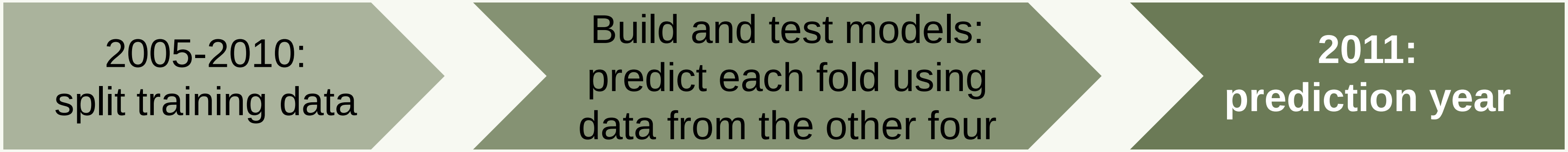
Great Lakes Beaches Modeled

The image part with relationship ID rId3 was not found in the file.

Wisconsin Department of Natural Resources

USGS Ohio Water Science Center

Building a PLS model – example for Edgewater, Ohio



2005-2010:
split training data

Build and test models:
predict each fold using
data from the other four

2011:
prediction year

divide the training data randomly into five equally-sized “folds”

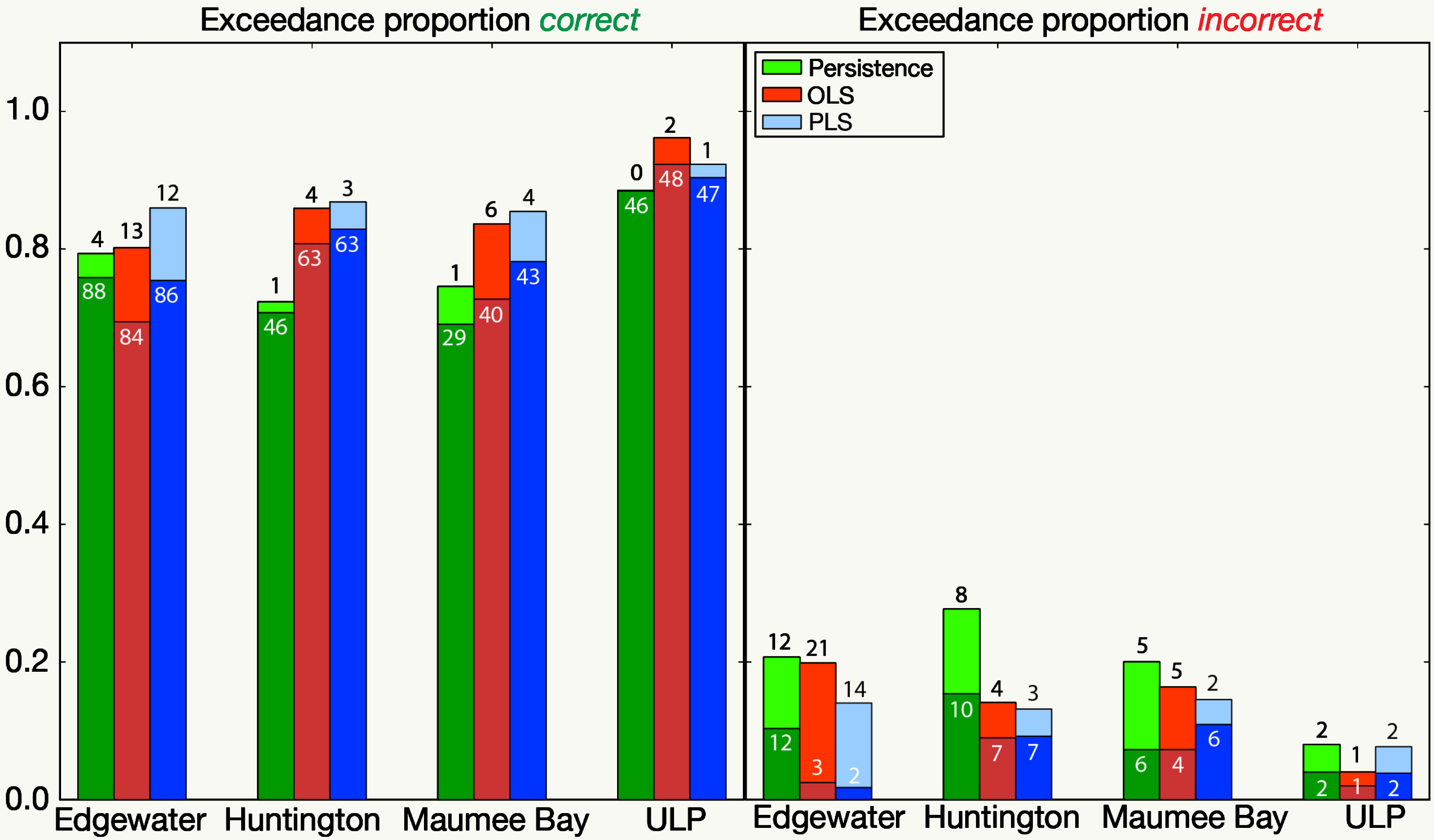
select a few candidate specificity limits (tuning)

compare model performance on the test folds

pick the specificity limit that had the best performance, and train a new model over all five folds

make and record predictions to manage beach and provide data going forward

2010 Model Performance

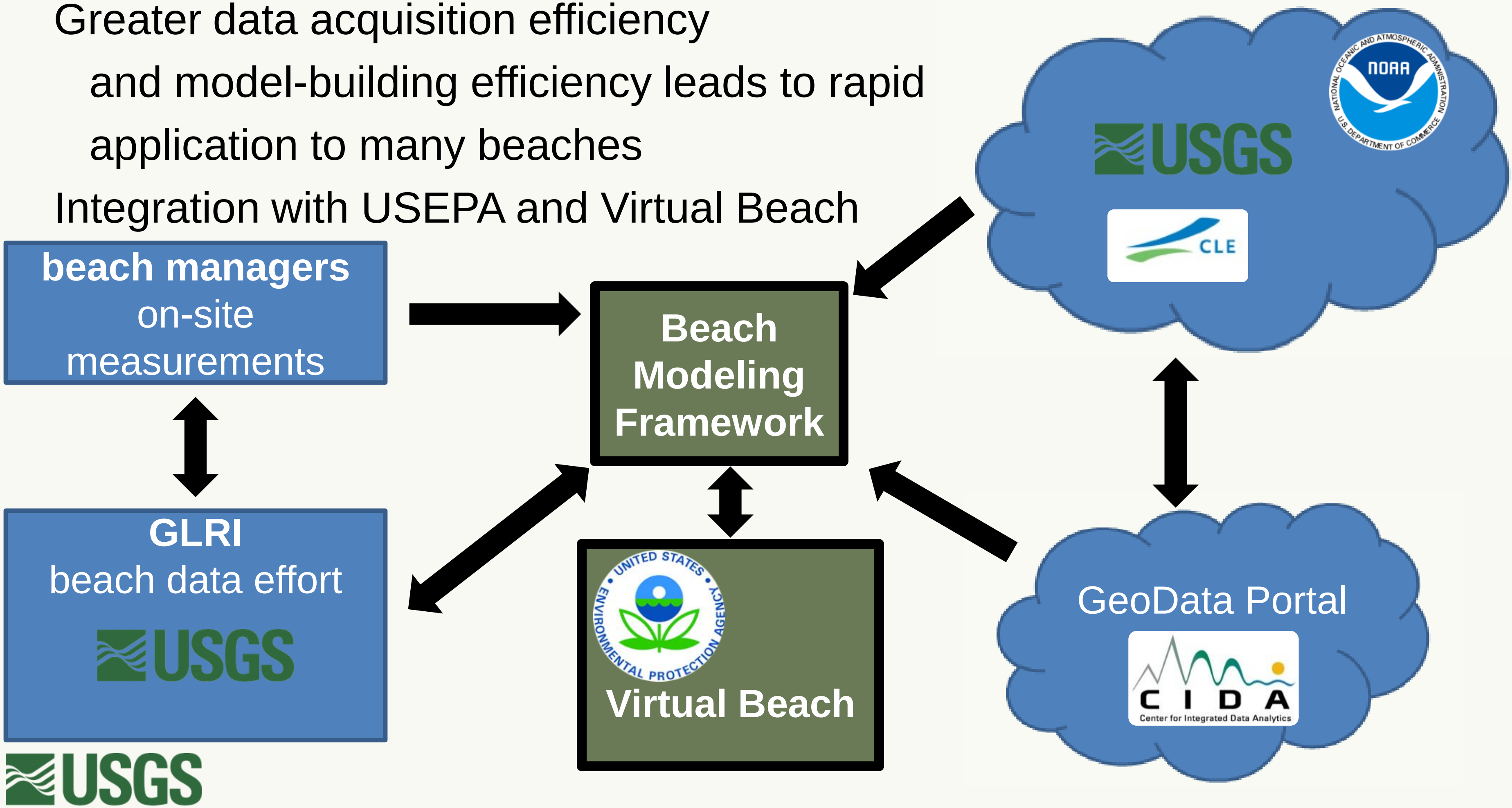


Next Steps

Data acquisition and connectedness

Greater data acquisition efficiency and model-building efficiency leads to rapid application to many beaches

Integration with USEPA and Virtual Beach



beach managers
on-site measurements

GLRI
beach data effort

Beach Modeling Framework

Virtual Beach

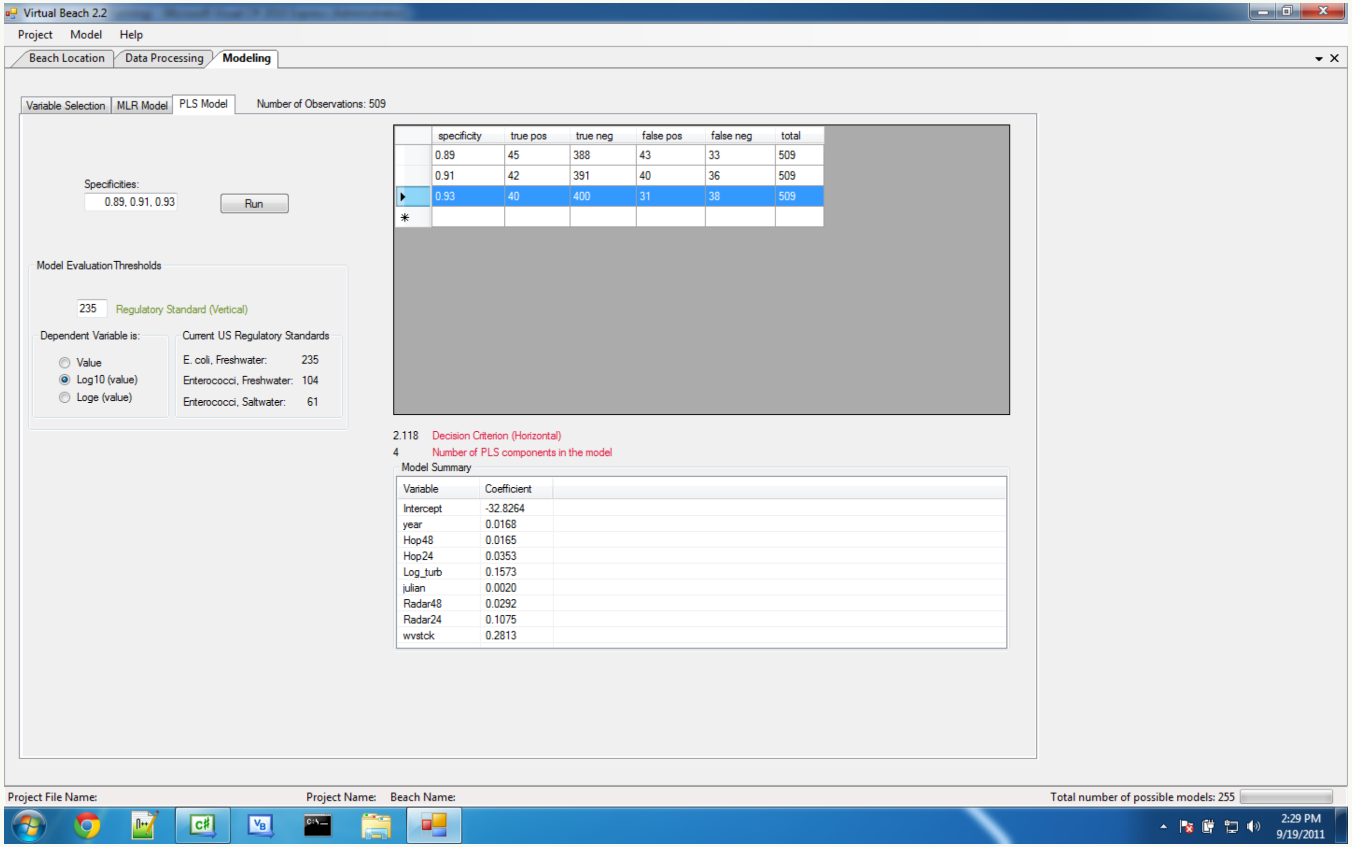
GeoData Portal

Next Steps

Virtual Beach

Well-known EPA software for predicting bacterial concentration

PLS regression to be included in version 3 (coming 2012)



Number of Observations: 509

Specificities: 0.89, 0.91, 0.93

	specificity	true pos	true neg	false pos	false neg	total
	0.89	45	388	43	33	509
	0.91	42	391	40	36	509
▶	0.93	40	400	31	38	509
*						

235 Regulatory Standard (Vertical)

Dependent Variable is:

○ Value
● Log10 (value)
○ Loge (value)

Current US Regulatory Standards

E. coli, Freshwater:	235
Enterococci, Freshwater:	104
Enterococci, Saltwater:	61

2.118 Decision Criterion (Horizontal)

4 Number of PLS components in the model

Variable	Coefficient
Intercept	-32.8264
year	0.0168
Hop48	0.0165
Hop24	0.0353
Log_turb	0.1573
julian	0.0020
Radar48	0.0292
Radar24	0.1075
wvstck	0.2813

Project File Name: Project Name: Beach Name: Total number of possible models: 255

2:29 PM
9/19/2011

Acknowledgements

OLS model results and collaboration:

Rob Darner

Donna Francy

Amie Brady

Dan Ziegler

Adam Mednick

Data provided by:

Cuyahoga County Board of Health

Northeast Ohio Regional Sewer District

Ozaukee County, Wisconsin

Funding provided by:

Ocean Research Priorities Plan (ORPP)

USEPA Great Lakes Restoration Initiative (GLRI)